

Forgetting the Words but Remembering the Meaning: Modeling Forgetting in a Verbal and Semantic Tag Recommender

Dominik Kowald
Know-Center
Graz University of Technology
Graz, Austria
dkowald@know-center.at

Christoph Trattner
Know-Center
Graz University of Technology
Graz, Austria
ctrattner@know-center.at

Paul Seitlinger
KTI
Graz University of Technology
Graz, Austria
paul.seitlinger@tugraz.at

Tobias Ley
Institute of Informatics
Tallinn University
Tallinn, Estonia
tley@tlu.ee

ABSTRACT

We assume that recommender systems are more successful, when they are based on a thorough understanding of how people process information. In the current paper we test this assumption in the context of social tagging systems. Cognitive research on how people assign tags has shown that they draw on two interconnected levels of knowledge in their memory: on a conceptual level of semantic fields or topics, and on a lexical level that turns patterns on the semantic level into words. Another strand of tagging research reveals a strong impact of time dependent forgetting on users' tag choices, such that recently used tags have a higher probability being reused than "older" tags. In this paper, we align both strands by implementing a computational theory of human memory that integrates the two-level conception and the process of forgetting in form of a tag recommender and test it in three large-scale social tagging datasets (drawn from BibSonomy, CiteULike and Flickr).

As expected, our results reveal a selective effect of time: forgetting is much more pronounced on the lexical level of tags. Second, an extensive evaluation based on this observation shows that a tag recommender interconnecting both levels and integrating time dependent forgetting on the lexical level results in high accuracy predictions and outperforms other well-established algorithms, such as Collaborative Filtering, Pairwise Interaction Tensor Factorization, FolkRank and two alternative time dependent approaches. We conclude that tag recommenders can benefit from going beyond the manifest level of word co-occurrences, and from including forgetting processes on the lexical level.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
RecSys'14, Silicon Valley, USA, October 06-10, 2014.
Copyright 2014 ACM xxx-x-xxxx-xxxx-x/xx/xx ...\$xx.xx.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—*Data mining*; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information filtering*

Keywords

personalized tag recommendations; time dependent recommender systems; Latent Dirichlet Allocation; LDA; human categorization; human memory model; BibSonomy; CiteULike; Flickr

1. INTRODUCTION

Many interactive systems are designed in a way that they mimic human behavior and thinking. For example, intelligent tutoring systems make inferences similar to teachers when they draw on knowledge of the learning domain, knowledge about the learner and knowledge about effective teaching strategies. Similarly, recommender systems based on Collaborative Filtering use information about socially similar individuals to recommend items, much in the same way as humans are influenced by similar peers when they make choices. An implicit assumption behind this seems to be that interactive systems should be better the closer they correspond to human behavior. Such assumption seems to be sensible because it is humans that interact with these systems and the systems often draw on data that humans have produced (such as in the case of the Collaborative Filtering approaches). It is therefore reasonable to assume that strategies that have evolved in humans over their individual or collective development are good models for interactive systems. However, the assumption that an interactive system should perform better the closer it mimics human behavior is not often tested directly.

In the current paper, we test this assumption in the context of a tag recommender algorithm. We draw on research that has explored how human memory is used in a dynamic and adaptive fashion to make sense of new information encountered in the environment. Sensemaking happens by dynamically forming ad-hoc categories that relate the new information with knowledge stored in the semantic memory (e.g., [4]). For instance, when reading an article about personalized recommendations, a novice has to figure out meaningful connections between previously distinct topics such as

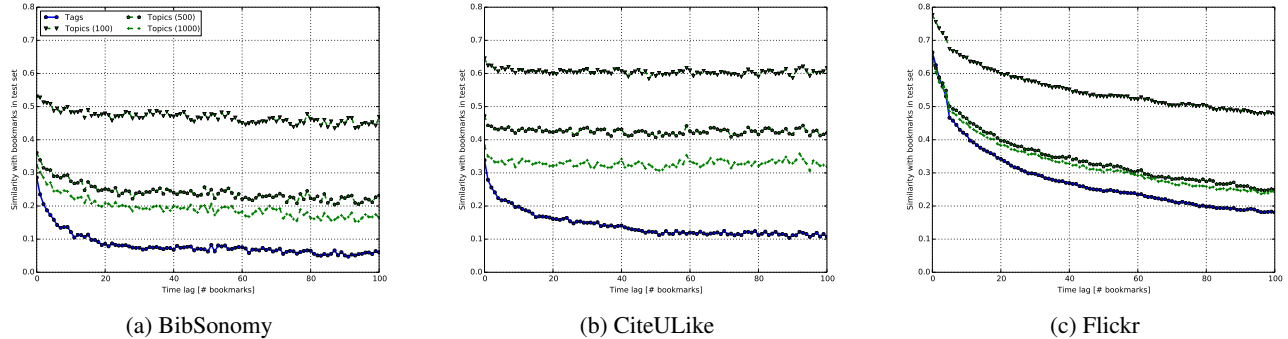


Figure 1: Interaction between time dependent forgetting and level of knowledge representation for BibSonomy, CiteULike and Flickr showing a more pronounced decline for tags than for topics (100, 500, 1000 LDA topics) (first research question).

cognition and information retrieval and hence, has to start developing an ad-hoc category about common features of both of them. When using a social tagging system in such a situation, people apply labels to their own resources which to some extent externalize this process of spontaneously generating ad-hoc categories [10]. Usually, a user describes a particular bookmark by a combination of about three to five tags verbalizing and associating aspects of different topics (e.g., “memory”, “retrieval”, “recommendations”, “collaborative filtering”).

In previous work, we have shown that this behavior can be well described by differentiating between two separate forms of information processing in human memory, a semantic process that generates and retrieves topics or gist traces, and a verbal process that generates verbatim word forms to describe the topics [27]. In this paper, we put an emphasis on another fundamental principle of human cognition to improve this model. According to Polyn et al. [24], memory traces including recently activated features contribute more strongly to retrieval than traces including features that have not been activated for a longer period of time. This relationship provides a natural account of what is called the recency effect in memory psychology (e.g., [2]). Obviously, things that happened a longer time ago tend to be forgotten and influence our current behavior less than things that have happened recently.

The purpose of this paper is twofold. First, we study the interaction between the effect of recency and the level of knowledge representation in human memory (semantic vs. verbal) in a social tagging system. In particular, we raise the question whether the impact of recency interacts with the level of knowledge representation, i.e., whether a time-dependent shift in the use of topics can be dissociated from a time-dependent shift in the use of particular tags (*first research question*). The second purpose, then, is to examine the question as to whether our tag recommender can be improved by integrating a time-dependent forgetting process and how this recommender performs in comparison to other well-established tag recommender algorithms (e.g., Collaborative Filtering, Pairwise Interaction Tensor Factorization and FolkRank), as well as two alternative time-dependent approach called GIRPTM [30] and BLL+C [17] (*second research question*).

The remainder of this paper is organized as follows. We begin with reviewing some of the work concerning recency in memory research and its current use in social tagging in Section 2. Then we describe our approach and the experimental setup of our extensive evaluation in Sections 3 and 4. We then present the results of this evaluation in terms of recommender quality in Section 5 and dis-

cuss related work in the field in Section 6. Finally, we conclude the paper by discussing our findings and future work in Section 7.

2. RECENCY IN MEMORY AND IN THE USE OF SOCIAL TAGGING

In previous work we have introduced 3Layers [27], a model for recommending tags that is inspired by cognitive-psychological research on categorizing and verbalizing objects (e.g., [10]). It consists of an input, a hidden and an output layer, where the hidden layer is built up by a semantic and an interconnected lexical matrix. The semantic matrix stores the topics of all bookmarks in the user’s personomy, calculated with Latent Dirichlet Allocation (LDA) [19], while the lexical matrix stores the tags of those bookmarks. In a first step of calculation, the LDA topics of a new bookmark, for which appropriate tags should be recommended, are represented in the input layer and compared with the semantic matrix of the hidden layer. In the course of this comparison, semantically relevant bookmarks of the user’s personomy become activated. The resulting pattern of activation across the semantic matrix is then applied to the lexical matrix to further activate and recommend those tags that belong to relevant bookmarks. In a final step, the activation pattern across the lexical matrix is summarized on the output layer in form of a vector representing a tag distribution that can be used to predict a substantial amount of variance in the user’s tagging behavior for the new bookmark.

We draw on Fuzzy Trace Theory (FTT; e.g., [5]) to make a prediction with respect to our first research question about a potentially differential impact of recency on semantic and lexical representations, i.e., on the usage of topics and tags, respectively. FTT differentiates between two distinct memory traces, a gist trace and a verbatim trace, which represent general semantic information of e.g., a read sentence and the sentence’s exact wording, respectively. These two types of memory traces share properties with our distinction between a semantic and a lexical matrix (see also Section 3). While vectors of the semantic matrix provide a formal account of each bookmark’s gist (its general semantic content), vectors of the lexical matrix correspond to a bookmark’s verbatim trace (explicit verbal information in form of assigned tags).

This assumption is also in line with Kintsch & Mangalath [16] who model gist traces of words by means of LDA topic vectors and explicit traces of words by means of word co-occurrence vectors. An empirically well-established assumption of FTT is that verbatim traces are much more prone to time-dependent forgetting than gist traces (e.g., [5]): while people tend to forget the exact wording,

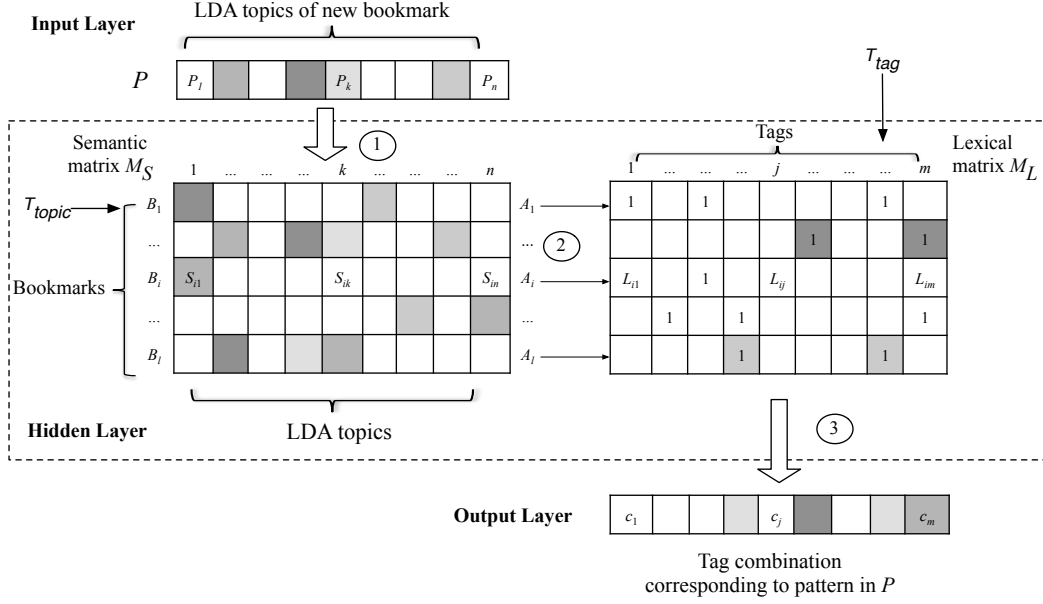


Figure 2: Schematic illustration of 3L showing the connections between the semantic matrix (M_S) encoding the LDA topics and the lexical matrix (M_L) encoding the tags. Furthermore, T_{topic} and T_{tag} schematically demonstrate how the time component is integrated in case of $3LT_{topic}$ and $3LT_{tag}$, respectively.

usually they can remember the gist of a sentence (or a bookmark). Taken together, we derived the hypothesis that a user’s verbatim traces (vectors in the lexical matrix that encode the user’s tags) are more strongly affected by time-dependent forgetting and therefore more variable over time than a user’s gist traces (vectors in the semantic matrix) that should be more similar to each other over time.

To test this hypothesis, we performed an empirical analysis in BibSonomy, CiteULike and Flickr (see Section 4.1). The topics for the resources of these datasets’ bookmarks were calculated using Latent Dirichlet Allocation (LDA) [19] (see Section 3) based on 100, 500 and 1000 latent topics in order to cover different levels of topic specialization. For each user we selected the most recent bookmark (i.e., the one from the test set with the largest timestamp, see also Section 4.2) and described the bookmark by means of two vectors: one encoding the bookmark’s LDA topic pattern (gist vector) and one encoding the tags assigned by the user (verbatim vector). Then, we searched for all the remaining bookmarks of the same user, described each of them by means of the two vectors and arranged them in a chronologically descending order. Next, we compared the gist and the verbatim vector of the most recent bookmark with the two corresponding vectors of all bookmarks in the user’s past by means of the cosine similarity measure.

The obtained results are represented in the three diagrams of Figure 1, plotting the average cosine similarities over all users against the time lags in days. For all three datasets we show these results for the last 100 bookmarks of tagging activity because in this range there are enough users available for each bookmark to calculate mean values reliably. The diagrams quite clearly reveal that – independent of the environment (BibSonomy, CiteULike or Flickr) – the similarity between the most recent bookmark and all other bookmarks decreases monotonically as a function of time lag. More importantly and as expected, the time dependent decline is more strongly pronounced for the verbatim vectors (encoding tag assignments) in contrast to the gist vectors (encoding LDA topics). Furthermore, we can see that the more LDA topics we use, the

more similar is the time dependent decline of the two vectors (tags vs. topics) to each other.

3. APPROACH

In this section we introduce two novel time-dependent tag recommender algorithms which model the process of forgetting on a semantic and lexical layer in a time-dependent manner. Based on our findings from the previous section, we assume that the factor of time plays a more critical role on the lexical layer than on the semantic layer. The approaches implemented in this section are based on a preliminary concept called 3Layers that was introduced in our previous work [27] to model semantic and lexical processes of tagging in social bookmarking systems.

Figure 2 schematically shows how 3Layers (3L) represents a user’s personomy within the hidden layer, which interconnects a semantic matrix, M_S (l bookmarks \times n LDA topics matrix), and a lexical matrix, M_L (l bookmarks \times m tags matrix). Thus, each bookmark of the user is represented by two associated vectors; by a vector of LDA topics $S_{i,k}$ stored in M_S and by a vector of tags $L_{i,j}$ stored in M_L . Similar to [20], we borrow a mechanism from MINERVA2, a computational theory of human categorization [13], to process the network constituted by the input, hidden and output layer. First, the LDA topics of the new resource to be tagged are represented on the input layer in form of a vector P with n features. Then, P is used as a cue to activate each bookmark (B_i) in M_S depending on the similarity (Sim_i) between both vectors, i.e., P and B_i . Similar to [20], we estimate Sim_i by calculating the cosine between the two vectors:

$$Sim_i = \frac{\sum_{k=1}^n P_k \times S_{i,k}}{\sqrt{\sum_{k=1}^n P_k^2} \times \sqrt{\sum_{k=1}^n S_{i,k}^2}} \quad (1)$$

To transform the resulting similarity values into activation values (A_i) and to further reduce the influence of bookmarks with low similarities, Sim_i is raised to the power of 3, i.e. $A_i = Sim_i^3$ (see

also [13]). Next, these activation values are propagated to M_L to activate tags that are associated with highly activated bookmarks on the semantic matrix M_S (circled numbers 2 and 3 in Figure 2). This is realized by the following equation that yields an activation value c_j for each of the m tags on the output layer:

$$c_j = \underbrace{\sum_{i=1}^l (L_{i,j} \times A_i)}_{3L} \quad (2)$$

To finally realize $3LT_{topic}$ and $3LT_{tag}$, we integrate a time component on the level of topics (hereinafter called T_{topic}) and on the level of tags (T_{tag}), respectively. Both recency components are calculated by the following equation that is based on the base-level learning (BLL) equation [2]:

$$BLL(t) = \ln((tmstp_{ref} - tmstp_t)^{-d}) \quad (3)$$

, where $tmstp_{ref}$ is the timestamp of the most recent bookmark of the user and $tmstp_t$ is the timestamp of the last occurrence of t , encoded as the topic in the case of T_{topic} or as the tag in the case of T_{tag} , in the user's bookmarks. The exponent d accounts for the power-law of forgetting and was set to 0.5 as suggested by Anderson et al. [1]. While $3LT_{topic}$ can be realized by using equation (4), $3LT_{tag}$ can be realized by using equation (5):

$$c_j = \underbrace{\sum_{i=1}^l (L_{i,j} \times \sum_{k=1}^n (S_{i,k} \times BLL(k)) \times A_i)}_{3LT_{topic}} \quad (4)$$

$$c_j = \underbrace{\sum_{i=1}^l (L_{i,j} \times BLL(j) \times A_i)}_{3LT_{tag}} \quad (5)$$

As described in [17], it is also important to take into account the tags that have been applied by other users to the target resource in the past in order to be able to also recommend new tags, i.e. tags that have not been used by the target user before. We do this by simply taking into account the most popular tags in the tag assignments of the resource Y_r (MP_r , i.e., $\arg \max_{t \in T} (|Y_r|)$) [14]. In order to combine c_j with MP_r , the following normalization method was used:

$$\|c_j\| = \frac{\exp(c_j)}{\sum_{i=1}^m \exp(c_i)} \quad (6)$$

Taken together, the list of recommended tags for a given user u and resource r is then calculated as

$$\tilde{T}(u, r) = \arg \max_{j \in T} (\beta \|c_j\| + (1 - \beta) \|Y_r\|) \quad (7)$$

, where β is used to inversely weight the two components. The results presented in Section 5 were calculated using $\beta = 0.5$, so giving the same weight to both components. The algorithms presented in this work are implemented in the Java programming language, are open-source software and can be exported online from our Github Repository¹ along with the test and training sets used for our experiments (see Section 4.1 and 4.2).

As outlined in Section 2, we calculated the semantic features of the resources in the bookmarks using *Latent Dirichlet Allocation* (LDA). In general, LDA is a probability model that helps to find

¹<https://github.com/learning-layers/TagRec/>

Dataset	$ B $	$ U $	$ R $	$ T $	$ TAS $
BibSonomy	400,983	5,488	346,444	103,503	1,479,970
CiteULike	379,068	8,322	352,343	138,091	1,751,347
Flickr	864,679	9,590	864,679	127,599	3,552,540

Table 1: Properties of the datasets, where $|B|$ is the number of bookmarks, $|U|$ the number of users, $|R|$ the number of resources, $|T|$ the number of tags and $|TAS|$ the number of tag assignments.

latent topics for documents where each topic is described by words in these documents [19]. This can be formalized as follows:

$$P(t_i|d) = \sum_{j=1}^Z P(t_i|z_i = j)P(z_i = j|d) \quad (8)$$

Here $P(t_i|d)$ is the probability of the i th word for a document d and $P(t_i|z_i = j)$ is the probability of t_i within the topic z_i . $P(z_i = j|d)$ is the probability of using a word from topic z_i in the document. The number of latent topics Z has to be chosen in advance, which defines the level of specialization of the topics. We calculated the semantic features for our datasets based on different numbers of LDA topics (100, 500 and 1000 - see also Section 5).

When using LDA in tagging environments, documents are resources which are described by tags. This means that resources in the bookmarks can also be represented with the topics identified by LDA based on the tag vectors of the resources (i.e., all the tags the users have assigned to the resource). These topics were then used as features in the semantic matrix M_S . We implemented LDA with Gibbs sampling using the Java framework Mallet².

4. EXPERIMENTAL SETUP

In this section we describe in detail the datasets, the evaluation methodology and the baseline algorithms used for our experiments.

4.1 Datasets

We used three well-known folksonomy datasets that are freely available for scientific purposes in order to conduct our study and for reasons of reproducibility. In this respect, we utilized datasets from the social bookmark and publication sharing system BibSonomy³ (2013-07-01), the reference management system CiteULike⁴ (2013-03-10) and the image sharing platform Flickr⁵ (2010-01-07) to evaluate our approach on both types of folksonomies, broad (BibSonomy and CiteULike; all users are allowed to annotate a particular resource) and narrow (Flickr; only the user who has uploaded a resource is allowed to tag it) ones [12]. We furthermore excluded all automatically generated tags from the datasets (e.g., *no-tag*, *bibtex-import*, etc.) and decapitalized all tags as suggested by related work in the field (e.g., [26]). In the case of CiteULike we randomly selected 10% and in the case of Flickr 3% of the user profiles for reasons of computational effort (see also [9])⁶. A p -core pruning approach was not applied in order to capture also the issue of cold-start users or items and to prevent a biased evaluation [7]. The statistics of our datasets can be found in Table 1.

²<http://mallet.cs.umass.edu/topics.php>

³<http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/>

⁴<http://www.citeulike.org/faq/data.adp>

⁵<http://www.tagora-project.eu/data/#flickrphotos>

⁶**Note:** We used the same dataset samples as in our previous work [17], except of CiteULike, where we used a smaller sample for reasons of computational effort in respect to the calculation of the LDA topics.

4.2 Evaluation Methodology

To evaluate our tag recommender approaches, we split the three datasets into training and test sets based on a leave-one-out hold-out method as proposed by related work in this field (e.g., [15]). Hence, for each user we selected her most recent bookmark (in time) and put it into the test set. The remaining bookmarks were then used for the training of the algorithms. This procedure simulates well a real-world environment because the tagging behavior of a user in the future is tried to be predicted based on the tagging behavior in the past. Furthermore, it is a standard procedure for the evaluation of time-based recommender systems [6].

To finally quantify the recommender quality and to benchmark our recommender against other tag recommendation approaches, a set of well-known metrics in information retrieval and recommender systems were used. In particular, we report Recall ($R@k$), Precision ($P@k$), Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) for $k = 10$ and F1-Score ($F_1@k$) for $k = 5$ recommended tags⁷ [21].

4.3 Baseline Algorithms

We compared the results of our approach to several “baseline” tag recommender algorithms. The algorithms were selected in respect to their popularity in the community, performance and novelty [3]. The most basic approach we utilized is the unpersonalized *MostPopular (MP)* algorithm that recommends for any user and any resource the same set of tags that is weighted by the frequency in all tag assignments [15]. A personalized extension of MP is the *MostPopular_{u,r} (MP_{u,r})* algorithm the suggests the most frequent tags in the tag assignments of the user (MP_u) and the resource (MP_r) [15]. Another simple and classic recommender approach is *Collaborative Filtering (CF)* which was adapted for tag recommendations by Marinho et al. [22]. Here the neighborhood of an user is formed based on the tag assignments in the user profile and the only variable parameter is the number of users k in this neighborhood. k has been set to 20 based on the work of Gemmell et al. [9].

Another approach we utilized is the well-known *FolkRank (FR)* algorithm which is an improvement of the *Adapted PageRank (APR)* approach [15]. FR adapts the PageRank algorithm in order to rank the nodes within the graph structure of a folksonomy[15] based on their importance in the network. Our implementation of APR and FR is based on the code and the settings of the open-source Java tag recommender framework provided by the University of Kassel⁸. A different popular and recent tag recommender mechanism is *Pairwise Interaction Tensor Factorization (PITF)* proposed by Rendle & Schmidt-Thieme [26]. It is an extension of *Factorization Machines (FM)* and explicitly models the pairwise interactions between users, resources and tags. The FM and PITF results presented in this paper were calculated using the open-source C++ tag recommender framework provided by the University of Konstanz⁹ with 256 factors as suggested in [26].

Finally, we also tried to benchmark against two time dependent approaches. The first one is the *GIRPTM* algorithm presented by Zhang et al. [30] which is based on the frequency and the temporal usage of a user’s tag assignments. The approach models the temporal tag usage with an exponential distribution based on the first- and last-time usage of the tags. The second time-dependent tag-recommender approach is the *Base-Level Learning Equation with*

⁷ $F_1@5$ was also used as the main performance metric in the ECML PKDD Discovery Challenge 2009: <http://www.kde.cs.uni-kassel.de/ws/dc09/>.

⁸<http://www.kde.cs.uni-kassel.de/code>

⁹<http://www.informatik.uni-konstanz.de/rendle/software/tag-recommender/>

	# Topics	Measure	3L	3LT _{topic}	3LT _{tag}
BibSonomy	100	$F_1@5$.197	.198	.204
		MRR	.152	.154	.161
		MAP	.201	.202	.212
	500	$F_1@5$.204	.205	.209
		MRR	.156	.158	.163
		MAP	.206	.208	.215
	1000	$F_1@5$.206	.207	.211
		MRR	.157	.158	.162
		MAP	.207	.208	.214
CiteULike	100	$F_1@5$.211	.212	.221
		MRR	.192	.194	.211
		MAP	.226	.228	.248
	500	$F_1@5$.218	.219	.225
		MRR	.196	.198	.211
		MAP	.232	.234	.250
	1000	$F_1@5$.232	.233	.238
		MRR	.199	.200	.212
		MAP	.235	.236	.250
Flickr	100	$F_1@5$.500	.507	.535
		MRR	.421	.429	.476
		MAP	.560	.571	.634
	500	$F_1@5$.564	.567	.582
		MRR	.443	.448	.476
		MAP	.591	.596	.635
	1000	$F_1@5$.568	.571	.585
		MRR	.450	.454	.477
		MAP	.599	.604	.636

Table 2: $F_1@5$, MRR and MAP values for BibSonomy, CiteULike and Flickr showing the performance of 3L and its time dependent extensions (3LT_{topic} and 3LT_{tag}) for 100, 500 and 1000 LDA topics (first research question).

Context (BLL+C) algorithm introduced in our previous work [17]. BLL+C is based on the ACT-R human memory theory by Anderson et al. [1] and uses a power-law distribution based on all tag usages to mimic the time dependent forgetting in tag applications. In both approaches the resource component is modeled by a simple most popular tags by resource mechanism, as it is also done in our 3Layers approaches.

5. RESULTS

In this section we present the evaluation of our two novel algorithms in two steps that correspond to our two research questions. In step 1, we compared the three 3Layers approaches (3L, 3LT_{topic} and 3LT_{tag}) with one another to examine our first research question of whether recency has a differential effect on topics and tags. Referring to our empirical analysis in Section 2, 3LT_{tag} should yield more accurate predictions than 3LT_{topic} and 3L.

The results in Table 2 are well in accordance with this assumption since - independent of the metric ($F_1@5$, MRR and MAP) and the number of LDA topics (100, 500, and 1000) applied - the difference between 3LT_{tag} and 3L appears to be larger than the one between 3LT_{topic} and 3L. Hence, a user’s gist traces (LDA topics) associated with the user’s bookmarks are less prone to “forgetting” than a user’s verbatim traces (tags associated with the bookmarks). Interestingly, this effect seems to be more strongly pronounced under the narrow folksonomy condition (Flickr) than under the broad folksonomy condition (BibSonomy and CiteULike).

Furthermore, Table 2 shows the performance of 3L, 3LT_{topic} and 3LT_{tag} for different numbers of LDA topics (100, 500 and 1000). In general these results reveal that all three approaches provide reasonable results for different levels of topic specialization and that

	$ B_{min} $	Measure	MP	LDA	MP_u	MP_r	$MP_{u,r}$	CF	APR	FR	FM	PITF	GIRPTM	BLL+C	3L	$3LT_{topic}$	$3LT_{tag}$
BibSonomy	-	$F_1@5$.013	.097	.152	.074	.192	.166	.175	.171	.122	.139	.197	.201	.206	.207	.211
		MRR	.008	.083	.114	.054	.148	.133	.149	.148	.097	.120	.152	.158	.157	.158	.162
		MAP	.009	.101	.148	.070	.194	.173	.193	.194	.120	.150	.200	.207	.207	.208	.214
	20	$F_1@5$.019	.142	.156	.078	.195	.204	.184	.197	.162	.163	.240	.249	.264	.269	.296 [°]
		MRR	.011	.129	.135	.059	.160	.175	.159	.171	.135	.137	.201	.216	.224	.227	.251 [°]
		MAP	.012	.152	.163	.074	.200	.219	.197	.214	.164	.166	.256	.275	.289	.291	.325 [°]
CiteULike	-	$F_1@5$.007	.068	.182	.033	.199	.157	.162	.160	.113	.130	.207	.215	.232	.233	.238 [°]
		MRR	.005	.065	.164	.024	.179	.168	.181	.181	.116	.149	.196	.205	.199	.200	.212 [°]
		MAP	.005	.073	.191	.029	.210	.196	.212	.212	.132	.169	.229	.241	.235	.236	.250 [°]
	20	$F_1@5$.008	.145	.228	.031	.237	.228	.221	.225	.193	.196	.282	.298	.331*	.334*	.353 [°]
		MRR	.006	.144	.225	.022	.233	.271	.237	.239	.201	.210	.321	.335	.312	.316	.367 [°]
		MAP	.006	.162	.258	.028	.269	.308	.273	.276	.229	.237	.369	.389	.369	.373	.430 [°]
Flickr	-	$F_1@5$.023	.169	.435	-	.435	.417	.328	.334	.297	.316	.509	.523	.568***	.571***	.585 [°]
		MRR	.023	.171	.360	-	.360	.436	.352	.355	.300	.333	.445	.466	.450	.454	.477 [°]
		MAP	.023	.205	.468	-	.468	.581	.453	.459	.384	.426	.590	.619	.599	.604	.636 [°]
	20	$F_1@5$.030	.190	.382	-	.382	.495	.322	.334	.309	.309	.534	.553	.610***	.616***	.643 [°]
		MRR	.028	.174	.322	-	.322	.473	.309	.317	.290	.289	.485	.508	.478	.485	.530 [°]
		MAP	.029	.215	.427	-	.427	.655	.405	.419	.378	.376	.664	.701	.661	.670	.732 [°]

Table 3: $F_1@5$, MRR and MAP values for all the users in the datasets (BibSonomy, CiteULike and Flickr) and for users with a minimum number of 20 bookmarks ($|B_{min}| = 20$) showing that our time dependent $3LT_{tag}$ approach outperforms current state-of-the-art algorithms (second research question). The symbols *, ** and *** indicate statistically significant differences based on a Wilcoxon Ranked Sum test between 3L, $3LT_{topic}$, $3LT_{tag}$ and BLL+C at α level .05, .01 and .001, respectively; °, °° and °°° indicate statistically significant differences between our two time dependent approaches $3LT_{topic}$, $3LT_{tag}$ and 3L at the same α levels.

the best accuracy results are reached with 1000 LDA topics¹⁰. The $F_1@5$, MRR and MAP values calculated for 1000 topics are also used in Table 3 for the second evaluation step that is described in the next paragraph.

In a second step, we contrasted the performance of our approaches, especially $3LT_{tag}$, with several state-of-the-art algorithms to address our second research question of whether 3L and its two extensions can be implemented in form of effective and efficient tag recommendation mechanisms. First, Table 3 reveals that all personalized recommendation mechanisms clearly outperform the unpersonalized MP approach, which simply takes into account the tag’s usage frequency independent of information about a particular user or resource.

Second and more important, 3L and its two extensions ($3LT_{topic}$ and $3LT_{tag}$) appear to reach higher accuracy estimates than the well-established mechanisms LDA, $MP_{u,r}$, CF, APR, FR, FM and PITF. From this we conclude that predicting tags in form of psychologically plausible steps of calculation that turn a user’s gist traces into words yields tag recommendations that correspond well to the user’s tagging behavior.

Third, also the two other time dependent algorithms (GIRPTM and BLL+C) outperform these state-of-the-art approaches that do not take the time component into account and in the case of BLL+C also reach higher estimates of accuracy than our 3L approach. However, this relationship between the two mechanisms dramatically changes if we enhance 3L by the recency component at the level of tags. Actually, $3LT_{tag}$ appears to outperform BLL+C in terms of all three measures and across all three datasets. Finally, as Figure 3 shows, a very similar pattern of results becomes apparent if the dif-

ferent approaches are evaluated by plotting recall against precision for $k = 1 - 10$ recommended tags.

To furthermore proof our assumption that memory processes play an important role in social tagging systems, we also performed an experiment where we looked at users that have bookmarked a minimum of $|B_{min}| = 20$ resources (see also [23]). We conducted this experiment by applying a post-filtering method, i.e., recommendations were still calculated on the whole folksonomy graph but accuracy estimates were calculated only on the basis of the filtered user profiles (= 780 users in the case of BibSonomy, 1,757 in the case of CiteULike and 4,420 for Flickr). The results of the experiment are also shown in Table 3. We observe that in general the accuracy estimates of all algorithms are increasing. Furthermore, we can see that the difference between $3LT_{tag}$ and the other algorithms (including BLL+C) gets substantially larger the more user “memory” (history) is used. These differences between $3LT_{tag}$ and BLL+C as well as between $3LT_{tag}$ and 3L proved to be statistically significant based on a Wilcoxon Rank Sum test across all accuracy metrics ($F_1@5$, MRR and MAP) and all three datasets (see Table 3).

6. RELATED WORK

In contrast to this study, previous research on tag recommender systems has taken a more pragmatist stance, typically ignoring cognitive psychological models that can help in explaining how people tag (as it was shown in this work). To date, the two following approaches have been established – folksonomy-based and content-based tag recommender approaches [21]. In our work we focus on folksonomy-based approaches.

The probably most prominent work in this context is the work of Hotho et al. [14] who introduced an algorithm called FolkRank (FR) that has established itself as the most prominent benchmarking tag recommender approach over the past few years. Subsequent and other popular works in this context are the studies of Jäschke et

¹⁰**NOTE:** We also performed experiments with more than 1000 LDA topics (e.g., 2000, 3000, ...). However, as also shown by related work (e.g., [19, 18]) this step did not help in increasing the performance of the LDA-based tag recommenders.

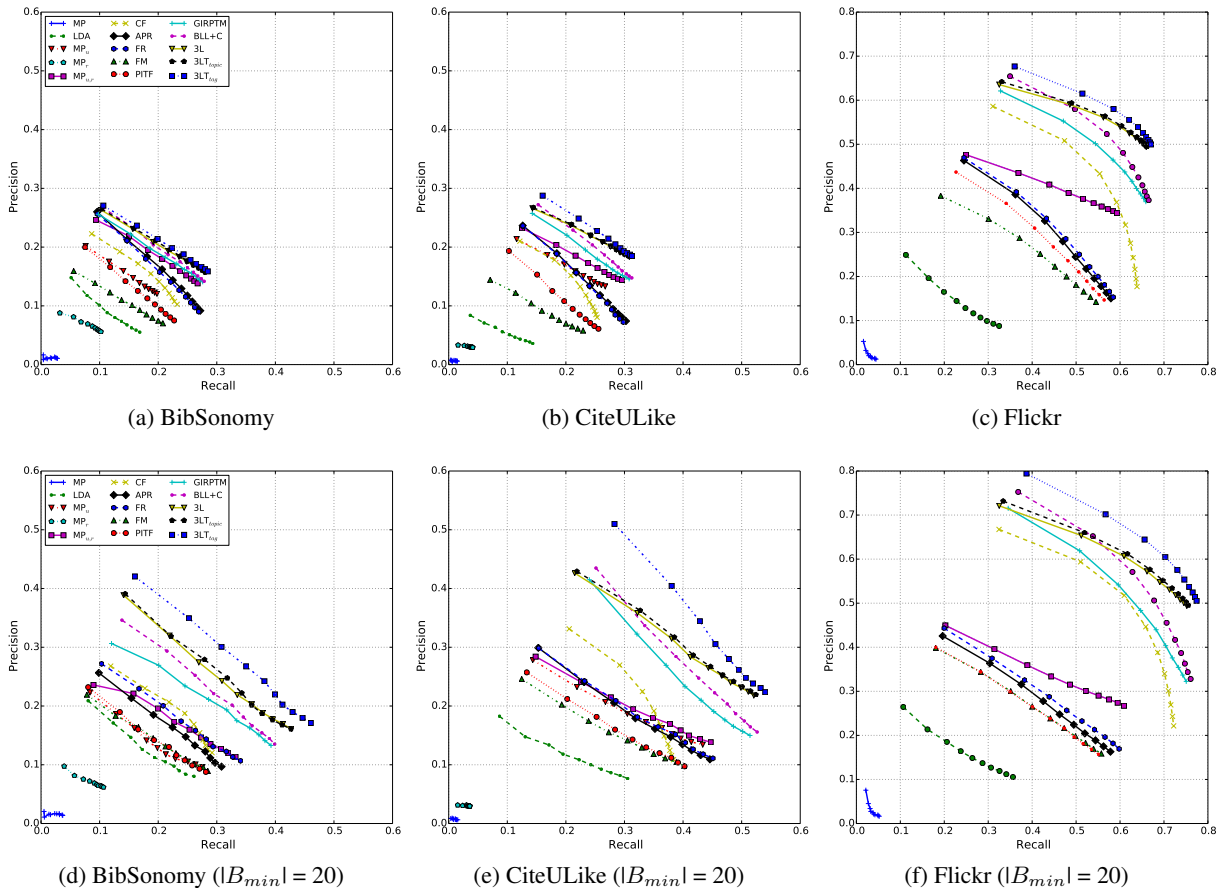


Figure 3: Recall/Precision plots for all the users in the datasets (BibSonomy, CiteULike and Flickr) and for users with a minimum number of 20 bookmarks ($|B_{min}| = 20$) showing the performance of the algorithms for 1 - 10 recommended tags (k).

al. [15] or Hamouda & Wanas [11] who introduced a set of Collaborative Filtering (CF) approaches for the problem of recommending tags to the user in a personalized manner. More recent and to some extent also well-know works are e.g., the studies of Rendle et al. [26], Krestel et al. [19], Rawashdeh et al. [25], Yin et al. [29] or Zhang et al. [30] who introduce a factorization model, a semantic model (based on LDA), a link prediction model or a time-based model to recommend tags to users (see Section 4.3).

Although the latter mentioned approaches perform more or less well in accurately predicting the users tags, all of them ignore well-established and long standing research from cognitive psychology on how humans process information. To bridge this gap we have recently introduced two simple and psychological plausible methods [27, 17] (= 3L and BLL+C) that are able (with limitations) to explain memory processes in social tagging systems. Based on these studies and new observations made in the current work, we were able to present a novel time-based tag recommender algorithm (= 3L_{tag}) in the end that significantly outperforms the state-of-the-art.

7. DISCUSSION AND CONCLUSION

In this study we have provided empirical evidence for an interaction between the level of knowledge representation (semantic vs. lexical) and time-based forgetting in the context of social tagging. Based on the analysis of three large-scale tagging datasets we con-

clude that - as expected - the gist traces of a user's personomy (the combination of LDA topics associated with the bookmarks) are more stable over time than the verbatim traces (the combination of associated tags). This pattern of results is well in accordance with research on human memory (e.g., [5]) suggesting that while people tend to forget surface details they keep quite robust memory traces of the general meaning underlying the experiences of the past (e.g., the meaning of read words). The interaction effect suggests that it is worthwhile to differentiate both time-based forgetting as well as level of knowledge representation in social tagging research.

Furthermore, the differential affect of forgetting on the two levels of processing has further substantiated the differences between tagging behavior on a semantic level of gist traces and a lexical level of verbatim traces [28]. This in turn is in line with cognitive research on social tagging (e.g., [8]) that suggests to consider a latent, semantic level (e.g., modeled in form of LDA topics) when trying to understand the variance in the statistical patterns on the manifest level of users' tagging behavior.

Finally, we have gathered further evidence for our assumption that interactive systems can be improved by basing them on a thorough understanding of how humans process information. We note in particular that integrating two fundamental principles of human information processing, time-based forgetting and differentiating into semantic and lexical processing, enhances the accuracy of tag predictions as compared to a situation when only one of the principles is considered. 3L, that is enhanced by forgetting on the lexical

level ($3LT_{tag}$), outperforms both the traditional 3L, as well as other well-established algorithms, such as CF, APR, FR, FM, PITF and the time-based GIRPTM. Furthermore, $3LT_{tag}$ also reaches higher levels of accuracy than BLL+C, the to-date leading time-based tag recommender approach, that was introduced in our previous work [17].

In future work, we plan to include our algorithms in a real on-line social tagging system (e.g., BibSonomy). Only in such setting is it possible to test the recommendation performance by looking at user acceptance. Because our approach is theory-driven, it is rather straightforward to transfer it to recommendations in other interactive systems and Web paradigms where semantic and lexical processing play a role (such as, for example, in Web curation). Generalization to other paradigms is another important benefit of driving recommender systems research by an understanding of human information processing on the Web.

Acknowledgments: The authors would like to thank Andreas Hotho and Denis Parra for many valuable comments on this work. This study is supported by the Know-Center, the EU funded project Learning Layers (Grant Agreement 318209) and the Austrian Science Fund (FWF): P 25593-G22.

8. REFERENCES

- [1] J. R. Anderson, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1050, 2004.
- [2] J. R. Anderson and L. J. Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.
- [3] L. Balby Marinho, A. Hotho, R. Jäschke, A. Nanopoulos, S. Rendle, L. Schmidt-Thieme, G. Stumme, and P. Symeonidis. *Recommender Systems for Social Tagging Systems*. SpringerBriefs in Electrical and Computer Engineering. Springer, Feb. 2012.
- [4] L. Barsalou. Situated simulation in the human conceptual system. *Language and cognitive processes*, 18(5-6):513–562, 2003.
- [5] C. Brainerd and V. Reyna. Recollective and nonrecollective recall. *Journal of memory and language*, 63(3):425–445, 2010.
- [6] P. G. Campos, F. Díez, and I. Cantador. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, pages 1–53, 2013.
- [7] S. Doerfel and R. Jäschke. An analysis of tag-recommender evaluation procedures. In *Proc. RecSys '13*, pages 343–346, New York, NY, USA, 2013. ACM.
- [8] W.-T. Fu and W. Dong. Collaborative indexing and knowledge exploration: A social learning model. *IEEE Intelligent Systems*, 27(1):39–46, 2012.
- [9] J. Gemmell, T. Schimoler, M. Ramezani, L. Christiansen, and B. Mobasher. Improving folkRank with item-based collaborative filtering. *Recommender Systems & the Social Web*, 2009.
- [10] R. J. Glushko, P. P. Maglio, T. Matlock, and L. W. Barsalou. Categorization in the wild. *Trends in cognitive sciences*, 12(4):129–135, 2008.
- [11] S. Hamouda and N. Wanas. Put-tag: personalized user-centric tag recommendation for social bookmarking systems. *Social network analysis and mining*, 1(4):377–385, 2011.
- [12] D. Helic, C. Körner, M. Granitzer, M. Strohmaier, and C. Trattner. Navigational efficiency of broad vs. narrow folksonomies. In *Proc. HT '12*, pages 63–72, New York, NY, USA, 2012. ACM.
- [13] D. L. Hintzman. Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, 16(2):96–101, 1984.
- [14] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *The semantic web: research and applications*, pages 411–426. Springer, 2006.
- [15] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007*, pages 506–514. Springer, 2007.
- [16] W. Kintsch and P. Mangalath. The construction of meaning. *Topics in Cognitive Science*, 3(2):346–370, 2011.
- [17] D. Kowald, P. Seitlinger, C. Trattner, and T. Ley. Long time no see: The probability of reusing tags as a function of frequency and recency. In *Proc. WWW '14*, New York, NY, USA, 2014. ACM.
- [18] R. Krestel and P. Fankhauser. Language models and topic models for personalizing tag recommendation. In *Proc. WI-IAT 2010*, volume 1, pages 82–89. IEEE, 2010.
- [19] R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In *Proc. RecSys 2009*, pages 61–68. ACM, 2009.
- [20] P. J. Kwantes. Using context to build semantics. *Psychonomic Bulletin & Review*, 12(4):703–710, 2005.
- [21] M. Lipczak. *Hybrid Tag Recommendation in Collaborative Tagging Systems*. PhD thesis, Dalhousie University, 2012.
- [22] L. B. Marinho and L. Schmidt-Thieme. Collaborative tag recommendations. In *Data Analysis, Machine Learning and Applications*, pages 533–540. Springer, 2008.
- [23] D. Parra-Santander and P. Brusilovsky. Improving collaborative filtering in social tagging systems for the recommendation of scientific articles. In *Proc. WI-IAT 2010*, volume 1, pages 136–142. IEEE, 2010.
- [24] S. M. Polyn, K. A. Norman, and M. J. Kahana. A context maintenance and retrieval model of organizational processes in free recall. *Psychological review*, 116(1):129, 2009.
- [25] M. Rawashdeh, H.-N. Kim, J. M. Alja'am, and A. El Saddik. Folksonomy link prediction based on a tripartite graph for tag recommendation. *Journal of Intelligent Information Systems*, pages 1–19, 2012.
- [26] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proc. WSDM 2010*, pages 81–90, New York, NY, USA, 2010. ACM.
- [27] P. Seitlinger, D. Kowald, C. Trattner, and T. Ley. Recommending tags with a model of human categorization. In *Proc. CIKM '13*, pages 2381–2386, New York, NY, USA, 2013. ACM.
- [28] P. Seitlinger and T. Ley. Implicit imitation in social tagging: familiarity and semantic reconstruction. In *Proc. CHI '12*, pages 1631–1640, New York, NY, USA, 2012. ACM.
- [29] D. Yin, L. Hong, and B. D. Davison. Exploiting session-like behaviors in tag prediction. In *Proc. WWW'2011*, pages 167–168. ACM, 2011.
- [30] L. Zhang, J. Tang, and M. Zhang. Integrating temporal usage pattern into personalized tag prediction. In *Web Technologies and Applications*, pages 354–365. Springer, 2012.